Hate Speech Detection Using Transformer Models

David Galati
M-BIT
University of Twente
Enschede, Netherlands
d.galati@student.utwente.nl

Cosmin Ghiauru
M-BIT
University of Twente
Enschede, Netherlands
c.ghiauru@student.utwente.nl

Hani Wakim
B-AT
University of Twente
Enschede, Netherlands
h.wakim@student.utwente.nl

Vitalii Fishchuk
M-BIT
University of Twente
Enschede, Netherlands
v.fishchuk@student.utwente.nl

Abstract

The rise of social media platforms has led to an increased need for effective hate speech detection to maintain a safe online environment. This study explores the performance of various machine learning models, comparing traditional approaches with state-of-the-art transformer-based models in the context of hate speech detection across different social media platforms. Using datasets from Reddit, Twitter, Facebook, and YouTube, we standardized and cleaned the data to ensure consistency and robustness. We evaluated traditional models including TF-IDF, Word2Vec trained from scratch, and Word2Vec pre-trained on Google News, as well as advanced transformer models like BERT, RoBERTa, ELECTRA, and GPT-3.5. Our findings reveal that pre-trained transformer models, particularly ELECTRA, outperform traditional methods, indicating the importance of advanced context-aware techniques in hate speech detection. This research underscores the necessity of tailored moderation strategies for different platforms and suggests future directions for improving cross-platform generalizability and model performance.

I. INTRODUCTION

Social media has been seeing a constant rise for almost two decades now, and so has people's need to express themselves online via pictures, likes, and personal opinions on various topics. While some might argue that everyone has the right to express themselves in any way they wish to, some lines must be drawn when people's means of expressivity are against general moral and ethical values. Hate speech is one of the most widely spread 'forbidden' practices across the internet, and social media platforms can descend into chaos if it is not kept under control and punished accordingly. According to recent studies, the prevalence of hate speech on social media platforms has reached alarming levels. The Anti-Defamation League's 2023 report on "The State of Online Hate and Harassment" found that 52% of Americans experienced harassment online, with 28% facing severe harassment including physical threats and stalking [6] This pervasive issue not only affects individual users but also threatens the integrity of online discourse and community cohesion.

A good example of this happening is the social media platform X (formerly known as Twitter), which has seen unprecedented levels of hate speech on all fronts due to its lack of hate speech moderation. A recent report by the Center for Countering Digital Hate (CCDH) [1] highlights the platform's failure to deal with the extreme levels of hate speech despite numerous reports. What can be deduced from it is that content moderation policies against hate speech are essential in order to create a safe and respectful online environment and ensuring that these policies can be enforced efficiently should be a priority for all social media platforms.

However, manual content moderation has proven insufficient to address the scale and complexity of online hate speech. The sheer volume of user-generated content, coupled with the nuanced and context-dependent nature of hate speech, necessitates the development of automated detection systems. Machine learning approaches offer the potential to efficiently process vast amounts of data and adapt to evolving forms of hate speech [2].

One of the main objectives of the study we have conducted is to evaluate the performance of various ML models by focusing on the difference between more traditional methods and newer transformer-based models in the context of hate speech detection across different social media platforms.

Our research uniquely contributes to the field by conducting an exploration and analysis of traditional machine learning techniques and state-of-the-art transformer-based models across datasets from multiple social media platforms. This cross-platform approach allows us to investigate the generalizability of hate speech detection models and uncover platform-specific challenges. By doing so, we aim to bridge the gap between academic research and practical implementation of hate speech detection systems in diverse online environments.

Thus, the research questions below have been formulated with an exploratory view in mind:

I. *What ML models are used for hate speech detection on social media?*
    a. *What are the main challenges in the context of hate speech detection?*
    b. *Which of the explored ML models performs the best in the context of hate-speech detection on social media?*
II. *Should social media hate speech detection ML models be trained only on the platform they will be used on, or can they take advantage from hate speech data across other platforms?*
III. *What insights can we find about hate speech across social media platforms?*

Answering these questions is vital for advancing the field of automated hate speech detection. By understanding the most effective approaches, the transferability of models across platforms, and the nuances of hate speech manifestation, we can develop more robust and adaptable systems for creating safer online spaces.

We employ a comprehensive approach, beginning with traditional ML techniques such as Fuzzy Logic, Artificial Neural Networks, and Decision Trees, and extending to advanced transformer models like BERT and RoBERTa and Electra. This study also explores the potential of hybrid models combining these approaches in the context of hate speech detection. The study utilizes diverse datasets from multiple social media platforms, enabling us to explore the generalizability of our models and identify platform-specific challenges.

The findings from this research have significant implications for content moderation practices on social media platforms. By improving the accuracy and efficiency of hate speech detection, we can contribute to creating safer and more inclusive online environments. Moreover, our cross-platform analysis provides insights that can inform policy decisions and help social media companies tailor their moderation strategies to the unique characteristics of their platforms. Ultimately, this research aims to contribute to the ongoing efforts to combat online hate speech and its detrimental effects on individuals and society at large.

For this study we used 3 datasets to come up with our results:

Dataset	Size	Reference
Reddit	13437 comments	[3]
Twitter	26954 comments	[4]
Facebook & Youtube	4644 comments	[5], [6]

II. LITERATURE REVIEW

Based on[7], [8], there are various ML techniques that can be used for hate-speech recognition. The main categories of techniques used are:

1. Fuzzy Logic (FL): Fuzzy Rule Based (FRB), Fuzzy Multi-Task Learning (FML)
2. Artificial Neural Networks (ANN): Recurrent NN, Convolutional NN, Multi-Layer Perceptron (MLP)
3. Deep Learning (DL): Long Short-Term Memory (LSTM), One dimensional CNN (CNN-1D)
4. Bayesian Networks (BN)
5. Genetic Algorithms (GA): Genetic Programming (GP)
6. Kernel Methods: Support Vector Machine (SVM)
7. Logistic Regression (LR):
8. Decision Trees (DT): J-48graft, Random Forest (RF)
9. Hybrid Methods: Methods above together with other ML techniques

On top of the 'traditional' methods above, Transformer-based models like BERT, roBERTa and LLaMA have have emerged and seem to have even more promising results due to their advanced context awareness.

Even though BERT came up in 2018, it is missing from most literature up to 2020. [7] mentions their potential superior results and use in the future, while others neglect it (and other transformers) completely in NLP hate-speech recognition writings, even in research done up to 2022 [8]. Even though literature up to two years ago hasn't been consistent with the latest practices regarding hate speech detection, we managed to find that in the scientific literature where transformer-based models were used, BERT and other bidirectional transformer models seem to be getting the best results in the context of single-model use [9], while the best overall results have

been observed in models that combine bidirectional transformer models with Neural Networks [10]

III. EXPERIMENTAL SETUP

1.1 DATA PROCESSING AND CLEANING

1.1.1 Data Loading and Initial Processing:

The datasets were initially loaded from different sources, including JSON and CSV files for Reddit and Twitter, respectively, and an Excel file for YouTube and Facebook data. Each dataset contains unique identifiers for the text content and associated labels indicating the presence of hate speech.

1.1.2 Data Standardization:

To facilitate analysis across multiple platforms, the data structures were standardized. Columns were renamed for consistency ('text' for the message content and 'label' for the classification of the message as hate speech or not). The labeling schemes were also standardized, converting various label formats to a binary system where '1' represents hate speech and '0' denotes non-hate speech.

1.1.3 Missing Values:

All datasets underwent a missing value analysis, with subsequent removal of any rows containing null entries to ensure the robustness of the training models used later in the analysis.

1.1.4 Text Data Cleaning:

The text data was cleaned to remove noise and standardize the input for modeling. This included the removal of:

- URLs to eliminate irrelevant web links.
- Usernames and retweet tags which could bias the hate speech detection.
- Special characters and emojis that do not contribute to the analysis of text sentiment.
- Control over punctuation retention was implemented, preserving question marks, exclamation points, and commas which can be significant in understanding the tone of a message.

Figure 1 Data Cleaning Steps

1.1.5 Outlier Detection and Removal:

Messages significantly deviating from the average message length (calculated as three standard deviations from the mean message length) were removed to prevent skewed analysis due to exceptionally long or short messages.

1.1.6 Duplicate and Conflicting Data Handling:

The datasets were purged of duplicates, particularly focusing on entries with conflicting labels to maintain data integrity and reliability. Only the entries with consistent labeling across duplicates were retained.

1.1.7 Balancing the Data:

Given the imbalance typically present in datasets where hate speech is a minority class, balancing techniques were applied. For datasets with a smaller proportion of hate speech, oversampling was used. Conversely, random undersampling was used for datasets where the non-hate speech examples were disproportionately high to equalize the class distribution.

1.1.8 Training and Test Split

Post-cleaning and balancing, each dataset was split into training and test sets. This split was stratified to maintain equal proportion of class labels in both sets, crucial for training unbiased and generalizable models.

1.2 TRADITIONAL APPROACH

Traditionally the approach to NLP classification tasks followed the same general idea: tokenize, create embeddings, train classifier, update embeddings/classifier weights. We follow this approach, using TF-IDF, W2Vec pretrained on Google News dataset and W2Vec trained from scratch by us. The detailed approach can be seen on the diagram below.

Figure 2 outlines the workflow for text classification using traditional machine learning approaches. It starts with tokenizing the text using different tokenizers (PunktSentenceTokenizer or TreebankWordTokenizer). The tokenized text is then used to create embeddings via TF-IDF, W2Vec, or pre-trained W2Vec models. These embeddings are averaged per text piece and fed into a trained Random Forest classifier for prediction. The predictions are evaluated against the true labels using metrics such as accuracy, precision, recall, and F1-score, with detailed metrics per class.

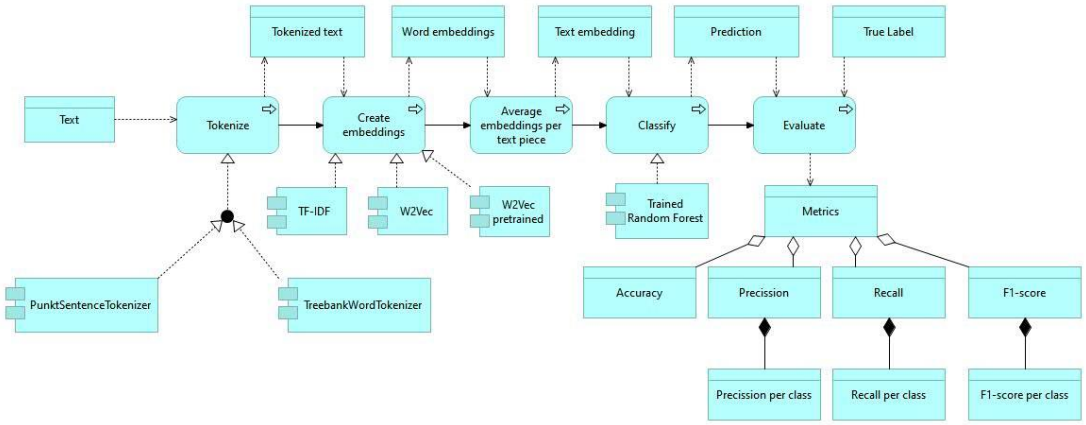

Figure 2 Traditional NLP Methodology

1.3 INNOVATIVE APPROACH

Following the traditional approach, we further explored recent developments in the field of hate-speech development: BERT, GPT, RoBERTA.

The model is used a simple classifier without fine-tuning/training, although two different system-level prompts are tested, simple and complex instructions. Fine-tuning wasn't performed due to limited resources and difficulty to select a subset of texts for fine-tuning, as feeding all of the training data to the GPT for fine-tuning is expensive. Same applies to few-shot learning, as it would require a careful selection of examples, which is out of the scope of this project. This is shown in Figure 3

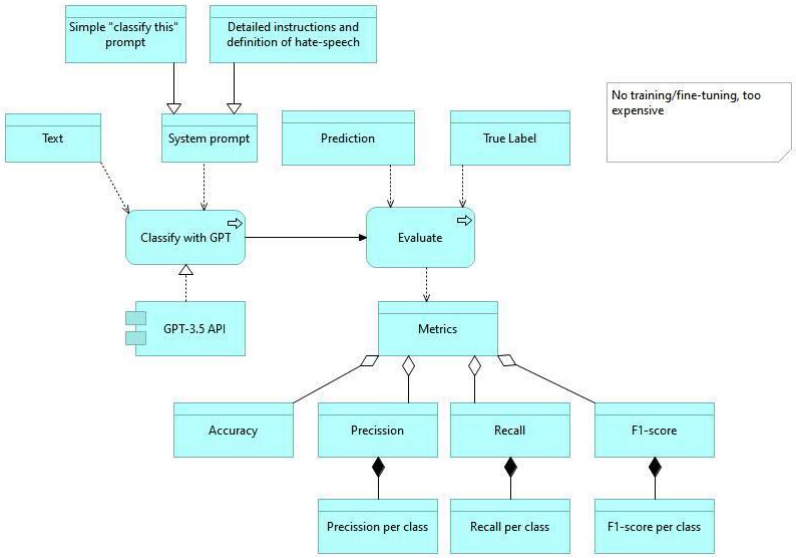

Figure 3 Transformer based methodology.

Three transformer-based models were trained and evaluated:

1. **BERT (Bidirectional Encoder Representations from Transformers):** A transformer-based model that uses a bidirectional training approach to capture context from both directions in the text.
2. **RoBERTa (Robustly optimized BERT approach):** An improved version of BERT that optimizes the pre-training phase.
3. **ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately):** A model trained to distinguish between real and fake tokens generated by a generator network.

For each model, cross-platform experiments were conducted:

- Training on one platform and testing on the other three.
- Evaluating performance using precision, recall, F1-score, and accuracy.

IV. RESULTS AND DISCUSSIONS

The results section aims to present and analyze the performance of various machine learning models applied to hate speech detection across multiple social media platforms. Initially, we conducted baseline experiments using traditional methods such as Word2Vec (both trained from scratch and pre-trained on Google News) and TF-IDF to establish a performance benchmark. These methods are well-known in natural language processing for their ability to create meaningful word embeddings and term representations.

However, as the field of NLP has evolved, transformer-based models have emerged as state-of-the-art techniques for various text classification tasks, including hate speech detection. These models, such as BERT, RoBERTa, and ELECTRA, leverage advanced contextual embeddings and deep learning architectures to achieve superior performance. In our study, we also evaluated the performance of GPT-3.5, both without training and with limited fine-tuning, to explore the capabilities of cutting-edge language models in this domain.

By comparing the results of traditional models with those of transformer-based models, we aim to highlight the advancements in hate speech detection capabilities. The findings from these experiments will provide insights into the effectiveness of different approaches and inform future research and practical implementations for maintaining safer online environments.

1.4 TRADITIONAL MODELS

Figure 4, Figure 5 and Figure 6 show subtle performance increase for w2vec pretrained in comparison with w2v and tf-idf, highlighting the importance of pretrained models and their ability to exercise better performance using pre-trained embeddings. W2Vec was not fine-tuned in this study due to practical limitations, instead, the superior performance of pre-trained word2vec motivated us to examine performance of fine-tuned BERT and other, more complex models.

Another finding is the close resemblance between results of pre-trained word2vec and tf-idf, shedding light on how tf-idf embeddings learned on a sufficiently large dataset capture similar semantics and word2vec pretrained on a large news corpus.

Notably, performance of the models differs based on the training dataset (see figures above), with typically the highest metrics obtained when training and testing on the same dataset. This proves that hate-speech is not uniform across various social media, presenting a unique opportunity of data enrichment through aggregation. The combined training dataset graph illustrates this by displaying superior performance across all the experiments and embedding methods. Furthermore, we have observed anomaly, where training on certain datasets shows significant boost to classification on specific test sets from other platforms across all embedding methods. For example, training the classifier on YouTube dataset shows excellent performance on Facebook test dataset, even better than when training on Facebook and testing on Facebook, and higher than results from testing on YouTube dataset (see the figures). This behavior is persistent across all embedding methods and cannot be attributed to the datasets size, as both YouTube and Facebook are approximately of similar size and label distribution. Further inquiry is required to establish the cause.

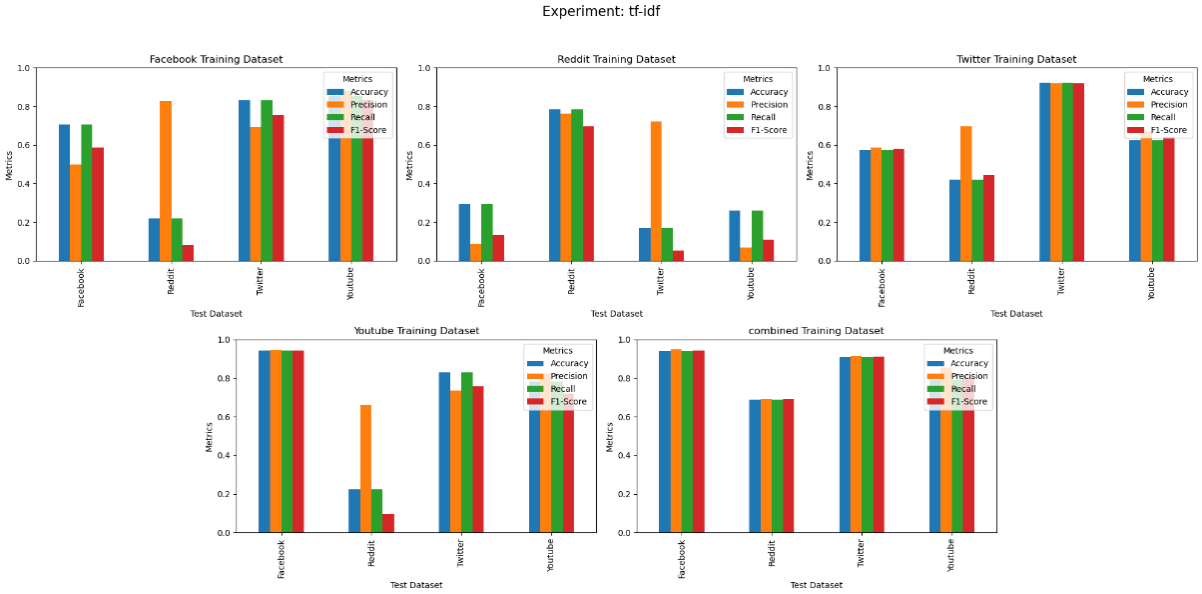

Figure 4

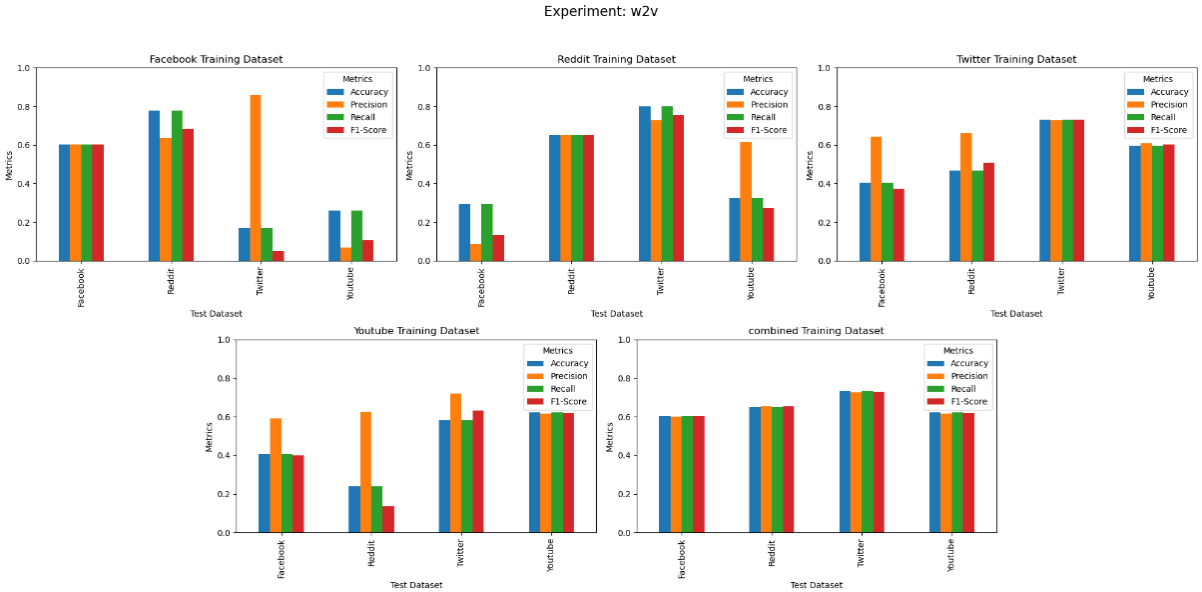

Figure 5

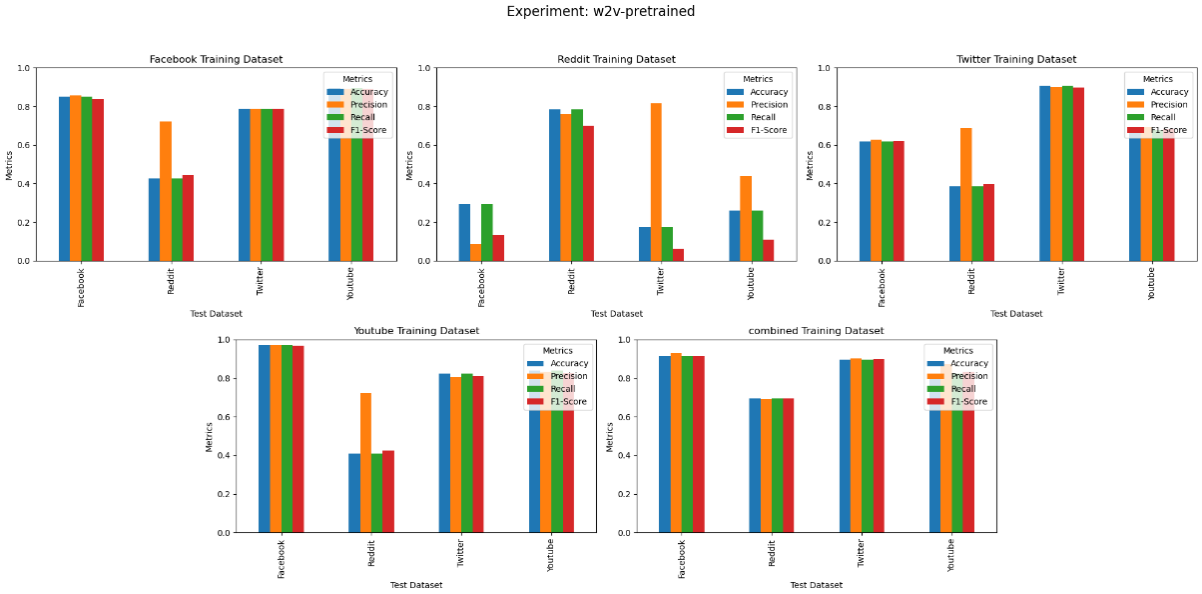

Figure 6

1.5 TRANSFORMER-BASED MODELS

1.5.1 GPT

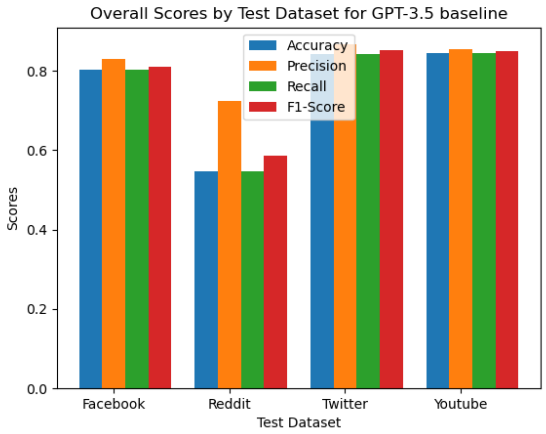

Figure 7

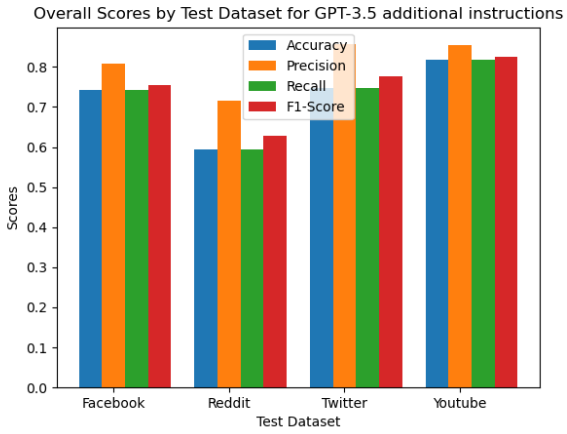

Figure 8

As seen in Figure 7 GPT has shown impressive performance in out-of-the-box configuration, however, due to limited interaction available with the GPT model (through black-box API), the performance is worse than even baseline models trained locally. We found the GPT model to have sub-par consistency and robustness, with same inputs producing different classifications in different runs under the same configuration. Prompt engineering had little to no impact on the results as well. As discussed in the literature above, GPT can outperform existing hate-speech classifiers as discussed by [11], however, this study found the model impractical in the real-world scenario. We have observed that training models locally provides similar/or better results at a

fraction of the costs, with full transparency and possibility for explainability, which is not present in the GPT API. GPT might be better suited for more complex tasks, but not for hate-speech classification.

1.5.2 Bert family and Electra

Now let's turn our attention to the performance of transformer-based models: BERT, RoBERTa, and ELECTRA. These models represent the state-of-the-art in natural language processing and have shown promising results in various classification tasks, including hate speech [11]. For each model these experiments have been conducted to extract metrics for helping in the discussion of model generalizability.

- Training on **Reddit**, Testing on Facebook, Twitter, and YouTube
- Training on **Facebook**, Testing on Reddit, Twitter, and YouTube
- Training on **Twitter**, Testing on Reddit, Facebook, and YouTube
- Training on **YouTube**, Testing on Reddit, Facebook, and Twitter

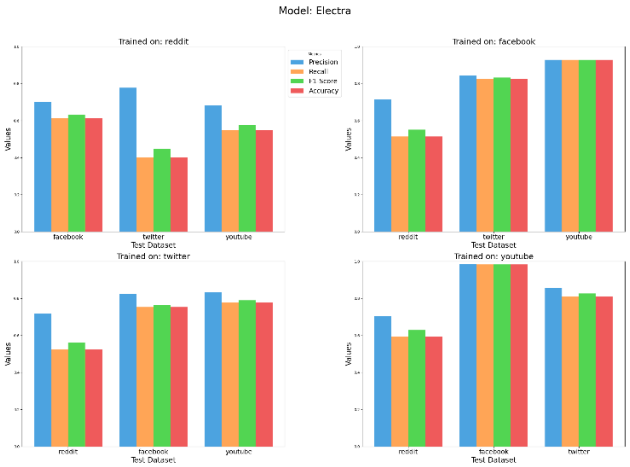

Figure 9

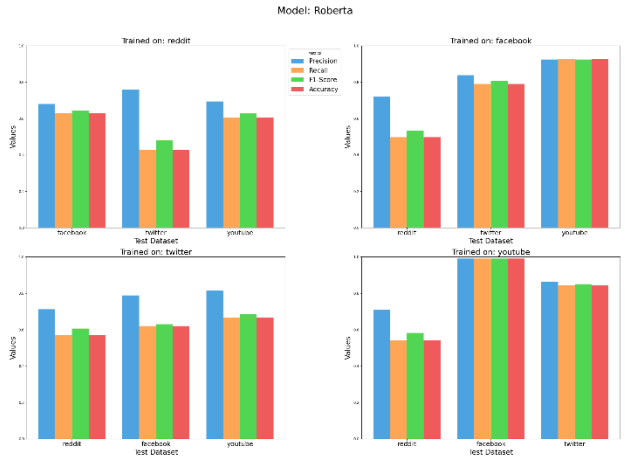

Figure 10

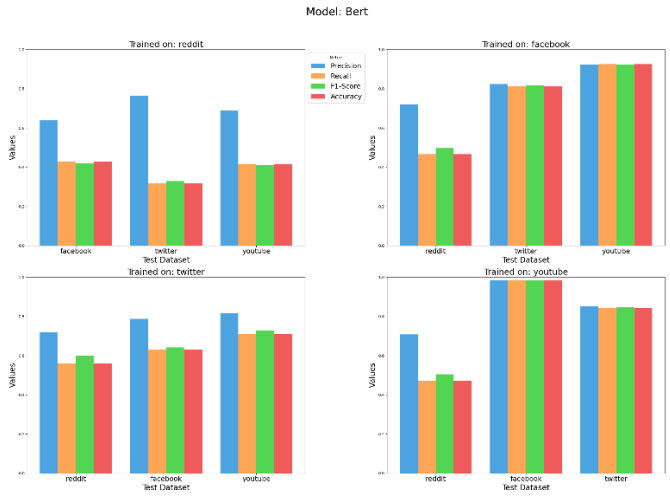

Figure 11

The experiments reveal interesting patterns in the cross-platform performance of BERT, RoBERTa, and ELECTRA. Notably ELECTRA consistently outperformed BERT and RoBERTa when trained on Reddit data and tested on other platforms. This aligns with [12] findings on ELECTRA's robust performance in transfer learning scenarios.

Compared to the traditional models discussed earlier, the transformer-based models demonstrated superior performance across all cross-platform tests. This superiority can be attributed to their ability to capture complex contextual information, as noted by Devlin et al. [13] in their work on BERT.

1.5.2.1 Specific Model Performance

BERT; The bidirectional nature of BERT, as described by [13], appears to contribute significantly to its performance in our cross-platform tests. Demonstrated strong performance due to its bidirectional nature, capturing context effectively from both directions.

ELECTRA; consistently outperformed BERT and RoBERTa when trained on Reddit data and tested on other platforms. This aligns with findings by Clark et al. [12] on ELECTRA's robust performance in transfer learning scenarios.

RoBERTa; Building upon BERT's architecture, RoBERTa showed marginal improvements over BERT in most scenarios, likely due to its optimized pre-training phase.

1.5.2.2 Domain Generalizability

Our results support the observations of [14] regarding the non-reversible nature of cross-domain performance in hate speech detection. For instance, while BERT performed well when trained on Reddit and tested on Twitter, the reverse did not hold true. Interestingly, training on the YouTube dataset yielded excellent performance on the Facebook test dataset, even surpassing the performance when trained and tested on Facebook data.

1.5.2.3 Challenges and Limitations

Despite strong overall performance, achieving consistently high precision and recall across all platforms remains challenging. This was particularly evident when models trained on one platform performed poorly on another, highlighting the difficulty in generalizing across different social media environments. highlighting the ongoing difficulty in hate speech detection noted by Malik et al. [8].

1.5.2.4 Implications of Findings

The findings suggest that while transformer-based models offer significant improvements over traditional approaches, the choice of model may need to be tailored to the specific platforms involved in hate speech detection. The strong performance of ELECTRA, in particular, points to the potential benefits of its discriminative pre-training approach.

1.5.2.5 Comparison with Previous Literature

Additionally, the results largely corroborate the findings of [12], [13] regarding the effectiveness of BERT and ELECTRA in classification tasks. However, it's observed that despite strong overall performance, achieving consistently high precision and recall across all platforms remains challenging. This was particularly evident when models trained on one platform performed poorly on another, highlighting the difficulty in generalizing across different social media environments.

1.5.2.6 Future Directions

Future research could explore fine-tuning strategies to improve cross-platform generalization, particularly for challenging platform combinations identified in our study. Additionally, investigating the impact of platform-specific language patterns on model performance could yield valuable insights for improving hate speech detection across diverse social media environments.

V. CONCLUSION

Our research explored the effectiveness of various machine learning models in detecting hate speech across multiple social media platforms, including Reddit, Twitter, Facebook, and YouTube. We conducted baseline experiments using traditional models such as TF-IDF, Word2Vec trained from scratch, and Word2Vec pre-trained on Google News. These were compared with state-of-the-art transformer-based models including BERT, RoBERTa, ELECTRA, and GPT-3.5.

The findings reveal that traditional models, while effective to a degree, are significantly outperformed by transformer-based models in hate speech detection. Transformer models demonstrated superior accuracy and robustness, leveraging advanced context-aware techniques. Specifically, BERT and RoBERTa provided substantial improvements over traditional approaches, and ELECTRA showed the highest performance due to its unique pre-training approach.

The results underscore the importance of utilizing advanced transformer-based models for hate speech detection. These models' ability to understand and process contextual information allows for more accurate classification, which is crucial for maintaining safe online environments. This has significant implications for data science and machine learning, as it highlights the necessity of adopting state-of-the-art models to tackle complex NLP tasks effectively.

Despite the promising results, the study faced limitations, particularly in the consistency and robustness of GPT-3.5 in a black-box setting. Fine-tuning and few-shot learning were not fully

explored due to resource constraints. Future research should focus on these areas to enhance the model's performance. Additionally, further investigation into cross-platform generalizability and the impact of platform-specific language patterns could provide valuable insights for improving hate speech detection across diverse social media environments.

The study found that transformer-based models significantly outperformed traditional models like TF-IDF and Word2Vec. Pre-trained models, especially those fine-tuned for specific tasks, demonstrated superior performance, highlighting the limitations of traditional models in handling complex linguistic nuances.

GPT-3.5, while powerful, showed sub-par performance in this study due to the constraints of a black-box API and the lack of fine-tuning. In contrast, BERT and RoBERTa, which were fine-tuned, provided better and more consistent results. This suggests that while GPT-3.5 has potential, its practical application in hate speech detection may be limited without extensive customization.

Both BERT and RoBERTa showed strong performance in hate speech detection, with RoBERTa marginally outperforming BERT in most scenarios. Their ability to capture bidirectional context proved crucial for accurate classification, confirming their superiority over traditional methods.

Our analysis revealed that models trained on combined datasets performed better across different platforms, suggesting that hate speech characteristics can be generalized to some extent. However, platform-specific nuances still play a significant role, and models tailored to individual platforms showed the highest accuracy.

The study found that Reddit was the most challenging platform for hate speech detection. This is likely due to the platform's diverse user base and the variety of content, which makes it harder to establish consistent patterns for hate speech. Future research should focus on developing more sophisticated models and techniques to address these challenges on Reddit.

VI. ACKNOWLEDGEMENT

We acknowledge the assistance provided by large language models such as ChatGPT and GitHub Copilot in this project. These tools were invaluable in generating content, refining writing, and providing coding support throughout the research process. Their contributions greatly enhanced the efficiency and quality of our work.

REFERENCES

[1] C. for C. D. H. Inc, "New report: X Content Moderation Failure." Center for Countering Digital Hate | CCDH, Sep. 2023. [Online]. Available: https://counterhate.com/research/twitter-x-continues-to-host-posts-reported-for-extreme-hate-speech/.

[2] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language." Mar. 2017. doi: 10.48550/arxiv.1703.04009.

[3] "reddit_comments." Accessed: May 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/ignaciorusso/reddit-comments/versions/1?resource=download&select=Reddit

[4] "Hate Speech and Offensive Language Dataset." Accessed: May 17, 2024. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

[5] J. Salminen *et al.*, "Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, Art. no. 1, Jun. 2018, doi: 10.1609/icwsm.v12i1.15028.

[6] "ICWSM18 - SALMINEN ET AL.xlsx," Dropbox. Accessed: Jul. 04, 2024. [Online]. Available: https://www.dropbox.com/scl/fi/wvh2hk4jfkpflc2s3z1bs/ICWSM18%20-%20SALMINEN%20ET%20AL.xlsx?dl=0&e=2&rlkey=z5sbmyd17azobihg096v65y26&st=n39xg8en

[7] F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions," *Computer Science Review*, vol. 38, p. 100311, Nov. 2020, doi: 10.1016/j.cosrev.2020.100311.

[8] P. Fortuna, M. Domínguez, L. Wanner, and Z. Talat, "Directions for NLP Practices Applied to Online Hate Speech Detection." Jan. 2022. doi: 10.18653/v1/2022.emnlp-main.809.

[9] S. González-Carvajal and E. C. Garrido-Merchán, "Comparing BERT against traditional machine learning text classification," *JCCE*, vol. 2, no. 4, pp. 352–356, Apr. 2023, doi: 10.47852/bonviewJCCE3202838.

[10] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media." Cornell University, Oct. 2019. doi: 10.48550/arxiv.1910.12574.

[11] N. Bauer, M. Preisig, and M. Volk, "Offensiveness, Hate, Emotion and GPT: Benchmarking GPT3.5 and GPT4 as Classifiers on Twitter-specific Datasets," in *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, R. Kumar, A. Kr. Ojha, S. Malmasi, B. R. Chakravarthi, B. Lahiri, S. Singh, and S. Ratan, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 126–133. Accessed: Jul. 04, 2024. [Online]. Available: https://aclanthology.org/2024.trac-1.14

[12] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," *International Conference on Learning Representations*, Apr. 2020.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North*, vol. 1, 2019, doi: 10.18653/v1/n19-1423.

[14] J. Malik, H. Qiao, A. van den Hengel, and G. Pang, “Deep Learning for Hate Speech Detection: A Comparative Study.” arxiv.org, Dec. 2023. Accessed: Jul. 04, 2024. [Online]. Available: https://arxiv.org/html/2202.09517v2